

A szövegtárak szókincsének összehasonlítása szótári címszójegyzék felhasználásával – neologizmusok és archaizmusok detektálása

Kiss Gábor¹, Kiss Márton¹

¹ TINTA Könyvkiadó, Budapest
{kissgabo, kissmarci}@tintakiado.hu

Kivonat: A Magyar Történeti Korpusz (TK) és a Magyar Nemzeti Szövegtár (MNSz) összehasonlítása egy más irányú lexikográfiai feladat elvégzésének melléktermékeként jött létre a TINTA Könyvkiadóban. Az elsődleges feladat az Értelmező szótár+ (ÉrtSz+ [1]) címszavainak gyakorisági mutatóval való ellátása volt. A gyakorisági mutatók meghatározásához felhasználtuk mindkét magyar szövegtárat. Az elsődleges feladat elvégzése során megvizsgáltuk, hogy az ÉrtSz+ 15.850 címszava előfordul-e, és ha igen, milyen gyakran a fenti két magyar szövegtárban külön-külön. A két korpuszból kinyert gyakorisági adatok segítségével (súlyozást is alkalmazva) állapítottuk meg az egyes címszók gyakorisági osztályát, azaz soroltuk be az 5 gyakorisági osztály valamelyikébe.

1 A vizsgálat előzménye

A Magyar Történeti Korpusz (TK) és a Magyar Nemzeti Szövegtár (MNSz) összehasonlítása egy más irányú lexikográfiai feladat elvégzésének melléktermékeként jött létre a TINTA Könyvkiadóban. (Mint ismeretes, a TK-t a Magyar Nagyszótár munkálatai során építették fel, és ez a korpusz 18., 19. és 20. századi magyar szövegrészleteket tartalmaz, míg az MNSz-t a 20. század végén keletkezett elsősorban sajtónyelvi, irodalmi szövegek alkotják.)

Az elsődleges feladat az Értelmező szótár+ (ÉrtSz+ [1]) címszavainak gyakorisági mutatóval való ellátása volt. A gyakorisági mutatók meghatározásához felhasználtuk mindkét magyar szövegtárat¹. Az utóbbi évtizedekben a nemzetközi szótárirodalomban az angol értelmező szótárak nyomán elterjedt a címszavak gyakoriságának jelölése. A magyar szótárirodalomban a Magyar értelmező kéziszótár (ÉKsz.² [2]) közli elsőként a címszavak gyakoriságát.

Az ÉrtSz+ sajátosan izgalmas szint képvisel a magyar, de a nemzetközi szótárkinálatban is, mert erőteljesen túllép az értelmező szótár szokásos funkcióin. Az ÉrtSz+ a szómagyarázó funkció mellett – mint a szótár az alcímében is jelzi – Értelmezések, példamondatok, szinonimák, ellentétek, szólások, közmondások, etimológiák, nyelv-

¹ Ezúton is köszönetet mondunk az MTA Nyelvtudományi Intézetének a két korpusz szokásos felhasználási módját meghaladó vizsgálat engedélyezéséhez.

használati tanácsok és fogalomkörü csoportok szerint is átfogó módon dolgozza fel címszóállományát. Sőt a szótár még a címszó gyakoriságát is feltünteti egy ötfokozatú skálán.

Mint említettük, az ÉrtSz+ címszavai gyakoriságának meghatározásához felhasználtuk a TK-t és az MNSz-t. A feladat elvégzése során megvizsgáltuk, hogy az ÉrtSz+ 15.850 címszava előfordul-e, és ha igen, milyen gyakran a fenti két magyar szövegkorpuszban külön-külön. A két korpuszból kinyert gyakorisági adatokat – leegyszerűsítve mondva – átlagoltuk, majd súlyozást is alkalmazva állapítottuk meg az egyes címszók gyakorisági osztályát, azaz soroltuk be az 5 gyakorisági osztály valamelyikébe. A kapott eredmény elemzését követően, a gyakorisági mutatót néhány esetben szubjektív nyelvérzékünk alapján módosítottuk.

2 A vizsgálat

Szövegkorpuszok összehasonlítására nincs általánosan elfogadott módszer. A matematikai logika nyelvére lefordítva: két nagy, de véges számú, ismétlődő diszkrét elemeket is tartalmazó halmazt kell összevetni. A halmazokban nemcsak az egyes elemek, azaz szavak megléte vagy hiánya a fontos, hanem az is jellemző, hogy egy-egy szó hány-szor fordul elő egy-egy szövegkorpuszban.

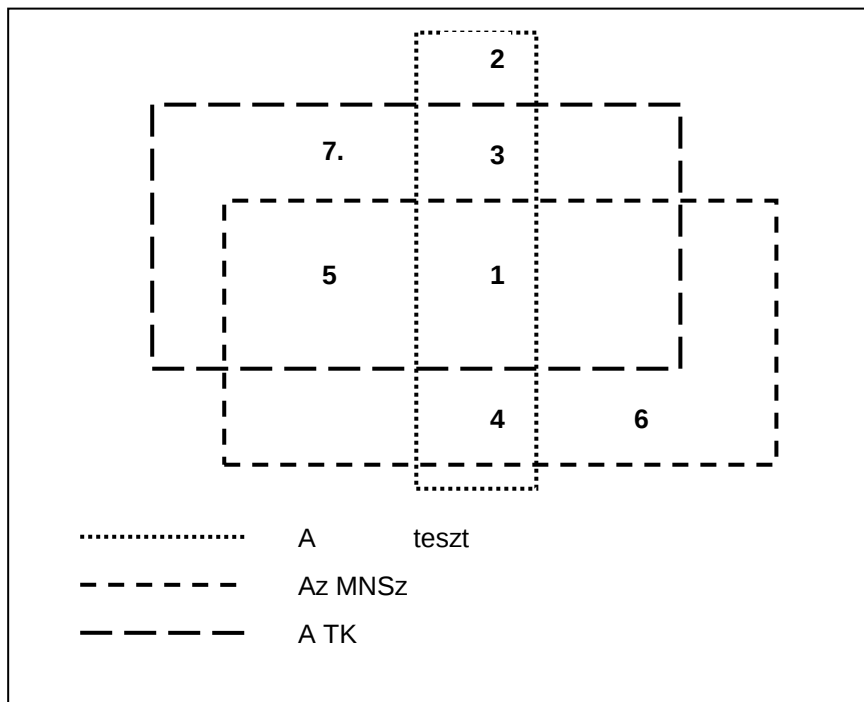
Mivel a gyakorlatban szinte lehetetlen két vizsgált nagy szövegkorpusz minden egyes szavának az összevetéséből keletkezett eredmény kiértékelése, az ÉrtSz+ címszavainak gyakorisági mutatójának a meghatározása megmutatott egy praktikusnak és helyesnek tűnő eljárást ahhoz, hogy két vagy több szövegkorpusz szókincsét miként lehet összevetni. Ugyanis a munkálat rávilágított arra, hogy egy jól megválasztott minta, vagy más néven tesztszólistának a szavait felhasználhatjuk a korpuszok szókincsének összevetésére, összehasonlítására. Természetesen ennek a tesztszólistának a hossza az összehasonlítandó korpuszok szókincséhez képest nem lehet se túl kicsi, se túl nagy. Az ÉrtSz+ címszavainak gyakorisági osztályokba sorolása során úgy tapasztaltuk, hogy a szótár 15.850 címszavának a két korpuszban való előfordulása és az előfordulások gyakorisága jól jellemzi és sajátosan leírja a fenti két korpuszt. Így az ÉrtSz+ címszójegyzékét joggal tekinthetjük vizsgálatunk tesztszólistájának.

Az ÉrtSz+ címszavait különös gonddal válogatták össze a szerkesztők. A címszólista összeállításáról így vallanak: „Milyen szavakat tartalmaz címszóként a szótár? Elsősorban az úgynevezett alapszókinsz elemeit, amelyeket a leggyakrabban és legtöbbször használunk, ezek között egyaránt vannak fogalomszók (...) és formaszók (...). Az alapszavak mellett sok a tantárgyi szakszó, olyanok, amelyekkel a diákok az irodalom-, a nyelv-, a történelem-, a matematika-, a fizikaórán találkozhatnak. (...) Megtalálható a szótárban több olyan régi szó, amelyek már csak irodalmi és történelmi szövegekben olvashatók (...). ezeken kívül vannak benne új, legtöbbször a modern technikával kapcsolatos szavak (...) Helyet kaptak a szótárban olyan szavak is, amelyeket Magyarországon nem használunk, de a határon túli magyarok életéhez hozzátartoznak.” [1: VIII].

Két korpusz (MNSz és TK) szókincsének (szavainak) és a tesztszólistaként használt ÉrtSz+ címszavainak összevetése során a következő elméleti esetek lehetségesek:

1. A tesztszólista egy adott szava mindkét korpuszban előfordul valahányszor.
2. A tesztszólista egy adott szava nem fordul elő egyik korpuszban sem.
3. A tesztszólista egy adott szava a TK korpuszban előfordul valahányszor, míg az MNSz-ben nem.
4. A tesztszólista egy adott szava az MNSz-ben előfordul valahányszor, míg a TK-ban nem.
5. A tesztszólistában nem szereplő szó mindkét korpuszban előfordul.
6. A tesztszólistában nem szereplő szó az MNSz-ben előfordul, de a TK-ban nem.
7. A tesztszólistában nem szereplő szó a TK-ban előfordul, de az MNSz-ben nem.

A Magyar Nemzeti Szövegtár (MNSz), a Magyar Történeti Korpusz (TK) és a tesztszólista szavainak lehetséges viszonya:



Vizsgálatunk körébe természetesen csak a 1–4. pontok tartoztak, az 5–7. pontok esetei kívül estek figyelmünkön. Mérésünket 2006 májusában végeztük. Ebben az időben az MNSz 111.746.000 szövegszó nagyságú, míg a TK 8.897.000 szövegszó terjedelmű volt. A mérés elején az ÉrtSz+ címszavai közül eltávolítottuk az 1-nél nagyobb homonimaindexszel ellátott címszavakat. A homonimák nélküli tesztszólista az eredetileg 15.810 címszóból álló lista helyett 15.010 szó hosszúságú lett. Ezt követően egy kis robotprogram lekérdezte a tesztszólistát alkotó 15.010 szó előfordulási gyakoriságát a két korpuszban. A MNSz és a TK korpuszok a szövegszavaik korábbi

morfológiai elemzése során – ha kis mértékben is – eltérő morfológiai elemzési technikát alkalmaztak, ezért az egyes esetekhez tartozó szavak számát 10-es értékre kerekítettük.

3 A vizsgálat eredményei

3.1 Hasonlósági mutatók

Számszerűsítve a következő eredményt kaptuk a teszt szólista 15.010 szava és a vizsgált két korpusz szavainak viszonyára:

1. A tesztszólista 15.010 szavából 14.290 szó előfordult mind a TK-ban, mind az MNSz-ben; 95,20%
2. A tesztszólista 15.010 szavából 45 szó nem fordult elő egyik szövegtörzsben sem; 0,30%
3. A tesztszólista 15.010 szavából 670 szó csak az MNSz-ben fordult elő, a TK-ban nem szerepelt; 4,46%
4. A tesztszólista 15.010 szavából 5 szó csak a TK-ban fordult elő, az MNSz-ben nem szerepelt; 0,03%.

Az adatokból látszik, hogy a két magyar szövegtörzs szókincse hasonló, jelentős átfedést mutat egymással. A tesztszólistának használt ÉrtSz+ címszójegyzéke a szótár jellegéből és funkciójából adódóan jobban illeszkedik a MNSz-hez, mint a TK-hoz.

3.2 Neologizmusok

Az MNSz jellegéből adódóan joggal feltételezhetjük, hogy azok a szavak, amelyek csak az MNSz-ben fordulnak elő, jól jellemzik a 20. század végét, illetve az ezredfordulót, és így ezek a magyar szókincs neologizmusai közé tartoznak. Feltételezésünk igazolására közreadjuk annak az 50 leggyakoribb szónak a listáját, amelyek az MNSz-ben előfordulnak, de a TK-ban nem találhatók meg (a szó melletti szám a szó MNSz-beli előfordulását mutatja).

Neologizmusok, az MNSz leggyakoribb korfeszítő szavai:

<i>közszolgálati</i> 6764	<i>környezetvédelem</i>	<i>foci</i> 2541
<i>munkáltató</i> 6263	3592	<i>bevásárlóközpont</i>
<i>honlap</i> 5383	<i>digitális</i> 3590	2476
<i>euró</i> 4515	<i>szia</i> 3475	<i>kosárlabda</i> 2323
<i>közterület</i> 4415	<i>drog</i> 3399	<i>világháló</i> 2294
<i>sportág</i> 4204	<i>parkoló</i> 3373	<i>társasház</i> 2234
<i>CD</i> 3890	<i>atomerőmű</i> 3217	<i>tömegközlekedés</i> 2051
<i>piacgazdaság</i> 3743	<i>elsőfokú</i> 3184	<i>informatika</i> 1993
<i>e-mail</i> 3728	<i>internet</i> 3057	<i>globalizáció</i> 1833
<i>szoftver</i> 3611	<i>közalkalmazott</i> 3008	<i>videó</i> 1726

<i>bróker</i> 1566	<i>rajzfilm</i> 955	<i>sci-fi</i> 761
<i>multi</i> 1378	<i>mobiltelefon</i> 912	<i>papírforma</i> 755
<i>természetvédelem</i>	<i>tévécsatorna</i> 873	<i>hardver</i> 735
1136	<i>interaktív</i> 860	<i>lakópark</i> 693
<i>AIDS</i> 1079	<i>kemping</i> 852	<i>akciófilm</i> 666
<i>telefax</i> 1060	<i>versenyszféra</i> 803	<i>sikerdíj</i> 636
<i>hobbi</i> 1055	<i>éllovas</i> 794	<i>bulvárlap</i> 635
<i>tizenéves</i> 970	<i>elektronika</i> 789	

3.3 Archaizmusok

Ezt követően felmerül a kérdés, hogy a fenti neologizmusokhoz hasonlóan a magyar nyelv archaizmusait is kigyűjthetjük-e a két korpusz szóanyagának az összevetéséből? Mivel vizsgálatunkban a TK-ban nem találtunk jelentős számban olyan szavakat, amelyek csak abban fordulnak elő, és nem találhatók meg a MNSz-ben – direkt módon nem gyűjthetünk archaizmusokat a TK-ból. Azonban adódik a gondolat, hogy talán archaizmusnak tekinthetők azok a szavak is, amelyek arányaiban jóval többször fordulnak elő a TK-ban, mint az MNSz-ben. A gondolat életrevalónak tűnik, bizonyításképpen alább közreadjuk az első 50 olyan szót, amelyek jelentősen többször fordulnak elő a TK-ban, mint az MNSz-ben (a szót követő szám az előfordulási arányt mutatja).

Archaizmusok, azok a szavak, amelyek arányaiban a TK-ban jelentősen többször fordulnak elő, mint az MNSz-ben:

<i>aprószenetek</i> 15	<i>hanga</i> 23	<i>midőn</i> 34
<i>asztag</i> 12	<i>hazámfia</i> 64	<i>nadály</i> 18
<i>beszély</i> 15	<i>hevenyében</i> 15	<i>okuláré</i> 11
<i>billikom</i> 45	<i>honfi</i> 28	<i>orozva</i> 27
<i>borong</i> 13	<i>honn</i> 47	<i>pacsuli</i> 20
<i>bőszült</i> 20	<i>horgany</i> 50	<i>patvar</i> 14
<i>burnót</i> 24	<i>ispán</i> 16	<i>sarjadék</i> 17
<i>csepű</i> 14	<i>játszi</i> 26	<i>sólya</i> 14
<i>csibuk</i> 28	<i>kebel</i> 16	<i>süveg</i> 13
<i>csigáz</i> 99	<i>kegyed</i> 25	<i>szövétnék</i> 11
<i>csöngettyű</i> 13	<i>komika</i> 30	<i>szüle</i> 18
<i>dicső</i> 13	<i>komorna</i> 24	<i>tekintetes</i> 18
<i>divatozik</i> 22	<i>kopja</i> 20	<i>téns</i> 54
<i>dragonyos</i> 16	<i>korhely</i> 13	<i>tragika</i> 21
<i>epeszt</i> 36	<i>ködmön</i> 13	<i>urambátyám</i> 20
<i>fejkötő</i> 20	<i>mál</i> 28	<i>vágta</i> 23
<i>findzsa</i> 21	<i>málé</i> 11	<i>várta</i> 109
<i>gondola</i> 16	<i>mente</i> 14	<i>vitézkötés</i> 13
<i>hajdankor</i> 23	<i>messzely</i> 28	

3.4 A magyar szókincs magja

Mind a két vizsgált korpuszban előfordul a tesztszólista szavainak 95,2%-a. Mint láttuk, nem mindegy azonban, hogy a közös szavak előfordulásának mi az aránya. A következőkben közreadunk mintaképpen 50 olyan gyakori szót, amelyek előfordulási aránya megegyezik vagy közel azonos mindkét szövegkorpuszban. Az 50 mintaszó mindegyikét az ÉrtSz+ leggyakoribb címszavai közül választottuk, azaz mindegyik az első gyakorisági kategóriába tartozik a szótár öt kategóriájából. A szavak mögött álló szám az MNSz-ben és a TK-ban lévő előfordulások aránya. Ha a szám nagyobb egy-nél, akkor arányaiban az MNSz-ben fordult elő többször a szó, ellenkező esetben a TK-ban gyakoribb.

Megállapíthatjuk, hogy ezek a szavak fontosak, a magyar szókincs középpontjában, magjában helyezkednek el, hiszen használatuk több mint két évszázadon át gyakori és állandó intenzitású a korpuszok adatai alapján.

Állandó intenzitású és gyakori szavak a magyar szókincs magjából:

<i>ad</i> 1,13	<i>két</i> 1,17	<i>rendel</i> 0,85
<i>csinál</i> 0,99	<i>kevés</i> 0,91	<i>rossz</i> 1,03
<i>elmege</i> 0,87	<i>könnyű</i> 0,83	<i>sok</i> 1,03
<i>esik</i> 1,15	<i>könyv</i> 1,03	<i>század</i> 0,89
<i>este</i> 0,92	<i>magas</i> 0,92	<i>széles</i> 0,89
<i>férfi</i> 1,17	<i>magyaráz</i> 1,00	<i>szeret</i> 1,10
<i>fiatal</i> 0,90	<i>mond</i> 0,83	<i>szó</i> 0,92
<i>gondol</i> 1,03	<i>nagyon</i> 1,18	<i>szolgál</i> 0,92
<i>győz</i> 0,86	<i>nap</i> 1,03	<i>tart</i> 1,18
<i>három</i> 0,99	<i>négy</i> 1,17	<i>tartozik</i> 1,04
<i>hat</i> (szn) 1,02	<i>nehéz</i> 0,93	<i>teremt</i> 1,15
<i>hisz</i> 0,85	<i>név</i> 1,20	<i>tud</i> 1,19
<i>hoz</i> 1,00	<i>nyár</i> 1,09	<i>út</i> 1,12
<i>idő</i> 1,13	<i>ok</i> 1,13	<i>város</i> 1,13
<i>ismer</i> 0,89	<i>olvas</i> 0,88	<i>vesz</i> 0,95
<i>jut</i> 1,06	<i>óra</i> 1,17	<i>víz</i> 0,87
<i>katona</i> 0,88	<i>orvos</i> 1,02	

4 Összegzés

Összegzésként elmondhatjuk, hogy újszerű vizsgálatunk bebizonyította, hogy két szövegkorpusz szókincsének összehasonlítása eredményesen elvégezhető egy kisebb alkalmas tesztszólista – akár egy szótár címszójegyzékének – segítségével.

Két különböző jellegű szövegkorpuszból hatékonyan gyűjthetők ki egy megfelelő tesztszólista segítségével a korpuszokra külön-külön is jellemző szavak, a mi esetünkben neologizmusok és archaizmusok. A kigyűjtött neologizmusok a 20. század végé-

nek, az ezredfordulónak a korfestői, míg az archaizmusok jól jellemzik a 18. század vége, illetve a 19. század magyar szókincsét.

Ugyanakkor megállapítható, hogy a két korpusz közös elemei közül azok, amelyek előfordulása magas és az előfordulások aránya közel azonos az egyes korpuszokban, a magyar szókincs magját képezik, így köznyelvi szótárak címszójegyzékének összeállításához eredményesen használhatók.

A két magyar szövegtörzs összehasonlítása során nyert eredmények rávilágítanak arra is, hogy több és különböző típusú magyar szövegtörzsről lenne szükség, mert a szövegtörzsek összevetésével jellegzetes csoportokat alkotó szavak gyűjthetők egybe többé-kevésbé automatikusan. A törzsekből kigyűjtött szócsoporthoz alkalmas kiindulásai lehetnek a magyar szókészlet különböző szempontú szótári munkálatainak.

Hivatkozások

1. Eöry V. (főszerk.): Értelmező szótár+. Értelmezések, példamondatok, szinonimák, ellentétek, szólások, közmondások, etimológiák, nyelvhasználati tanácsok és fogalomkörök csoportok. TINTA Könyvkiadó, Budapest (2007)
2. Pusztai F. (főszerk.): Magyar értelmező kéziszótár². Akadémiai Kiadó, Budapest (2003)